Beyond Test Scores: A Better Way to Measure School Quality

Book · December 2017

DOI: 10.4159/9780674981157

1 author:

Jack Schneider
University of Massachusetts Lowell
32 PUBLICATIONS **216** CITATIONS

Beyond Test Scores

A Better Way to Measure School Quality

JACK SCHNEIDER

Harvard University Press

Cambridge, Massachusetts and London, England 2017

—-1
—0
—+1

Introduction

THE IDEA that I would have to move was ridiculous, but that's what the numbers kept suggesting.

In the spring of 2013, the *Boston Globe* created an online tool called the Dreamtown Finder, which proposed to help people choose a place to live. The tool prompted users to rank values in six categories, and then produced a customized list of Massachusetts cities and towns—ostensibly ordered by fit. Naturally, one of the categories was schools.

Like many people, I rated the "schools" category as very important. I did the same for "fun," "location," and "housing cost." I enjoy living in a diverse neighborhood, so I toggled the "people like me" bar down. I was neutral on the remaining variable—"hipster."

The site told me to pack up my family and move—a quarter-mile west, into Cambridge. I clicked the "back" button, determined to stay in Somerville, where I currently reside and am quite happy. I altered my rankings, and resubmitted, but I couldn't make it work. That is, not until I dropped my concern with school quality down to zero.

—-1
—0
—+1

The Somerville schools certainly aren't perfect. Parents, teachers, and principals all have wish lists and complaints. However, if you spend enough time in any school district in America, you will learn that every community has its own set of priorities and concerns. People care enough about education to want more from it.

Yet few people in Somerville seemed unhappy with the schools. Like Americans in general, residents of the city tend to express relatively high levels of confidence in the schools their children attend—a surprising fact given the doom-and-gloom rhetoric that dominates headlines about the nation's educational woes.[1] From what I could see, their general satisfaction seemed warranted. Still a relative newcomer to the city, I had yet to visit every school, but I did know several of them, and I had already begun to work closely with the school across the street from our house—a K–8 elementary school that my daughter now attends. We were also, at the time, preparing to enroll our daughter in the district's early childhood center—a diverse school with what seemed to us a solid mix of traits. The faculty was warm and caring, children seemed happy, the building was well kept and decorated with student artwork, and students appeared engaged in activity as we walked through on a tour.

Of course, judging a school from the outside is incredibly difficult. In lieu of decent information, parents often rely on imperfect measures of quality—a new computer lab, for instance, an array of impressive-sounding classes, or a list of prestigious colleges attended by graduates. No one knows this better than the leaders of private schools, which must sell their clients on the idea that education—free to all children in the United States—should actually come with a hefty price tag. Go on a private school tour, and you will see lots of bells and whistles, but you will learn less than you imagine about what your child's actual experience at that school will be.

My wife is a teacher and the daughter of a retired administrator. I am a former high school teacher who left the classroom to become a professor of education; much of my research is in school quality. As such, we thought of ourselves as relatively savvy. Still, we wondered: Why would anyone think the Somerville schools are bad?

There were a few reasons.

Most obviously, the district's raw standardized test scores were relatively low.[2] That, certainly, could be explained. The district serves a very diverse range of students who come from a wide variety of economic, racial, and ethnic backgrounds—many from low-income and minority families. As an extensive body of research indicates, there is a strong connection between family background and standardized test scores, so it made sense that anyone who saw test scores as an indicator of school quality might be misled about the quality of Somerville's schools.[3]

Many parents also tend to use race as a proxy for school quality. Knowing that students of color have long been denied equal educational opportunities, middle-class white parents often shy away from schools with large concentrations of black and brown students.[4] In so doing, they exacerbate segregation, take their high levels of capital elsewhere, and ensure that people like them continue to avoid schools with large populations of color.

I hoped that the *Globe* wasn't using raw test scores to measure school quality, and though I was relatively certain they weren't using race as a variable, I was curious enough that I dug into the methodology.

As it turned out, schools were being ranked by SAT scores and teacher-student ratios. The former, as research has borne out, is more strongly correlated with socioeconomic status than it is with college performance.[5] In response, some colleges are dropping the requirement that prospective students sit for the test.[6] The latter measure, beyond

being generally misleading—it counts the number of adults in the building rather than the average class size—is often not meaningfully different across schools. Consequently, it can lead to perceived differences where there may be none.

These were bad measures of school quality. But I wasn't particularly surprised. Most available measures of school quality are weak, and user-friendly data interpreted by third parties tend to be even weaker.

I wasn't in a position to offer much more than criticism. Nevertheless, I began using my newly minted Twitter account to go after the *Globe*. "Tools like this," I tweeted, "could be powerful if the methodology weren't so simplistic." I included the link to the *Globe* website.

It wasn't much in terms of public engagement. But to my surprise, it led to several interesting conversations, and eventually produced a response from the designer of the tool, who contacted me directly. His question, in essence, was a challenge: Do you think you can do better?

Two weeks later we were sitting in my office at the College of the Holy Cross, a small liberal arts college forty miles west of Boston. I suggested a few measures I thought were better gauges of school quality, as well as some ways of adjusting for the different kinds of populations various schools work with. We looked at the statewide data available to us, and though they were limited, we agreed that we could certainly do better than what had been included in the Dreamtown Finder.

Using educational research to inform our decisions, we picked several measures that captured more about school quality. And recognizing the fact that rating schools is an inherently subjective enterprise, we designed a model that allowed parents to customize rankings based on personal values—values that, in our version of the tool, included academic growth, school resources, college readiness, school culture, and diversity. Rather than assume a single vision of a good school, our tool asked users to distribute an allotment of one hundred points, in five-point increments, according to their own

-1—
0—
+1—

personal values. Thus, a parent who values the diversity of the student population and also wants to assure that his child will be college-ready at the end of high school could weight these two variables more heavily than the others. Alternatively, a parent concerned first and foremost about academic growth could allocate all of her points to that category.

The tool had some strengths, perhaps chief among them its user-friendliness. We labeled variables using accessible terminology and offered brief descriptions of our measures. Although in several cases we used aggregate variables to simplify user interpretation, each variable was accompanied by a brief description of its importance and the methodology for measurement. For those interested, a much longer rationale, including research citations, was provided on a separate linked page. In short, if people wanted to use the page to learn about educational data, they could. We were proud of that.

The second great strength of our tool was its multivariate nature. Though still new in education, balanced scorecards are commonplace in the field of performance management.[7] And for good reason. In any complex enterprise, goals will be multiple and will not always overlap; and education is an extremely complex enterprise. States and districts have increasingly come to recognize this fact over the past few years, and in the future the balanced scorecard will be standard in education as well. However, it seemed important to us to send a message, loudly and clearly, that there are many ways to measure school quality.

Our tool's third strength was that it allowed users to personalize their weightings—and hence their rankings—thereby avoiding a single, rank-ordered list of schools and any commensurate stigmatization. To put it another way, because you might rate some aspects more highly than I do, your list of "top schools" will look different from my list. That's a good thing, first, because it disrupts the message that there is a clear hierarchy of schools, and second, because a "good"

—-1
—0
—+1

school is often good for a particular family with a particular set of values. Being able to tailor measures of school quality to those values is important in challenging the notion that there is one single model of a successful school. As I continue to believe, supporting this diversity of values and needs is an essential asset of any school system interested in serving a broad and diverse constituency.

In sum, our tool, which the *Globe* called the Dreamschool Finder, sent several important messages: that a lot goes into creating a good school, that test scores reflect only a fraction of what we want schools to do, and that not all people want the very same things in the very same amounts. It also opened the door to the possibility that we could celebrate schools for their different strengths—allowing users to view schools through the lens of "fit" rather than through a mono-modal lens of "one best school." Our tool also provided more easy-to-digest information that, we hoped, would drive high-quality decision making—among parents, community members, and educators.

Still, it was far from perfect.

First, we were severely limited by the data available. Relying on the state for our measures, we were forced to use imperfect proxies for a number of categories. Thus, despite our focus on moving *beyond* standardized test scores, we still relied heavily on scores from the Massachusetts state standardized test—the MCAS—for indicators of learning. Additionally, for dozens of important variables there were simply no data, and we lacked the resources to collect those data ourselves. We had to pretend that factors such as how safe students feel, or how strong student-teacher relationships are, didn't matter, even though we knew that they did.

Second, our data were produced from measures implemented *at most* twice yearly. Student attitudes, knowledge, and behavior change daily. Teachers have good days and bad days, and schools themselves transform over time. Yet current data collection procedures capture

only single moments in time. In an ideal world, we would want to figure out how to capture multiple snapshots across the school year—in order to produce more accurate data, as well as to produce a picture of change over time.

Third, our categories for school quality were educated guesses at best. To produce a truly effective measure of school quality, we would have needed to determine the full range of outcomes valued by parents, educators, and the public. Lacking the necessary time and resources to do that, we relied on our existing knowledge. If we were to do the job really well, we would not only want to study polling data but also to conduct some of our own research—through polling, surveying, and focus groups.

Imperfect though the tool was, however, it did generate some interesting conversations.

First, it challenged conventional thinking about school quality. If you placed equal weights on all of the variables for high schools in Massachusetts, for instance, Somerville High came in at number fifteen. To some people, this was a shock. When the *Globe* released the results of the 2013 MCAS test, Somerville High ranked 271st out of 354 schools in English and 262nd in math. But that, of course, was based on raw test scores—unadjusted for any student background variables—which often reflect more about family income and parental education than they do about school quality. Suddenly people were forced to confront the possibility that schools they knew to be "bad" might turn out to be otherwise. Alternatively, many schools assumed to be "good" because of their high test scores might actually have some things to work on.

Second, it led to conversations about how to improve the tool. I readily acknowledged the limits of the Dreamschool Finder and was content to have tinkered around the margins. What I didn't expect, however, was that I would be asked repeatedly how to make it better.

—-1
—0
—+1

Although I initially brushed that question aside, I found myself thinking more and more about it. What if we could collect data on anything? What would we go after?

Those conversations eventually led to meetings with Somerville mayor Joe Curtatone and then-superintendent Tony Pierantozzi. Joe had just started his sixth term as mayor of what the *Boston Globe Magazine* called the best-run city in Massachusetts. He had already secured several new subway stops for the city, brokered some smart development projects, established Somerville as a home to the arts, and strengthened an already very livable city; but he wanted to do more to support education than just increase the size of the budget. Tony was approaching what would be his last full year as superintendent, having managed the city's schools for nearly a decade. Both were willing to experiment, as was Tony's successor, Mary Skipper. All felt that the schools were on the verge of something special, and all of them felt that existing test scores failed to capture the real quality of education in the city. As Tony put it, "We're better than the data."

The idea was to collect more information about the performance of Somerville public schools, but it soon became clear to all of us that a better data system would do more than collect and disseminate new information. It would be a challenge to the existing test-based accountability system.

For years, parents and educators have been pushing back against the singular focus on standardized test scores in measuring school quality. Each time, however, they have been met by the same reply: "What do you propose to do instead?" After all, stakeholders in public education are desperate for information. How can policy leaders govern—in an already massive and complex system—if they don't have measures of success in their tool kits? How can parents advocate for their children? How can communities and their allies advocate for the vulnerable? As Tony Pierantozzi observed, the emphasis on data

collection, problematic though it may be, was a response to a "pretty miserable track-record of dealing with poor students, minority students, and English language learners . . . it was embarrassing."

Efforts to measure schools, in short, were not going to go away, but it seemed undeniable that test-based measurement was having a troubling impact on schools. Consider, for instance, how states use test scores to hold schools accountable. In most states, schools are responsible for raising the test scores of all students across a variety of demographic subgroups. If they do not, the state intervenes by imposing sanctions, penalties, and eventually school shutdowns. Yet, because students from low-income and minority families are likely to score lower on standardized tests, their schools are far more likely to be stigmatized by state intervention, or closed down, regardless of almost every other aspect of school performance. Such stigmas and penalties create churn in school staff, as teachers flee or are fired. They drive away parents with options. And they send a message to students that they are on a dead-end track.[8]

State accountability measures also have a second problematic impact. Because schools are held accountable for a narrow set of scores—generally on math and reading tests in grades three through eight, as well as one year of high school—school leaders have responded rationally: by narrowing the curriculum. Arts, history, science, health, and other aspects of a diverse curriculum have been cut back dramatically. Emphasis on test-aligned math and English instruction has been ratcheted up. Furthermore, even teachers outside those content areas have been asked to focus on them as much as possible. As a result, teachers are increasingly unhappy with their profession, and many are deciding to leave. According to a 2012 MetLife survey, only 39 percent of teachers indicated that they were "very satisfied" with their jobs.[9]

Students are unhappy, too. They are bombarded with benchmark tests, practice tests, diagnostic tests, and the high-stakes tests

—-1
—0
—+1

themselves—and all for what? For many, school feels increasingly irrelevant and uninspiring. New evidence also suggests that upswings in test scores, though they may be associated with greater acquisition of content knowledge, may not be associated with cognitive growth.[10]

To be fair, states have tried to improve the way they measure school quality. Prompted by the Every Student Succeeds Act—the successor to the much-reviled No Child Left Behind law—many states have incorporated measures of "nonacademic factors" into their accountability systems. In Massachusetts, for instance, the Department of Elementary and Secondary Education (ESE) generated a list of several dozen indicators "suggested by external stakeholders and ESE staff" that might help round out the state's picture of K–12 schools.[11] Nevertheless, test scores remain the coin of the realm for the state. By extension, test scores continue to play an outsized role in how parents and the public think about schools.

All of this means that districts—and particularly urban districts like Somerville, which often have lower raw standardized test scores because of their demographic makeup—should be seeking to build better measures that more fairly capture academic performance, and more fully reflect the range of qualities that characterize good schools.

Our project, consequently, had the potential not merely to overhaul how schools are measured in Somerville but also to stand as a model for other districts. It might be a response to the question "What do you propose to do instead?"

In the intervening months, I began thinking more about the project, and particularly about who might help us tackle the first step—building a new framework for evaluating school quality. One person who jumped to mind was Rebecca Jacobsen, a professor at Michigan State University and an expert on how the public values education. Along with Richard Rothstein and Tamara Wilder, she had written a book about what better measures of school quality might

look like—*Grading Education: Getting Accountability Right*—and had conducted research into how data on school quality are presented to parents. Rebecca expressed a strong interest in the project, and we soon began reviewing polling data on what Americans want their schools to do.

I also began looking for a survey design expert. After all, if we were going to build a new framework for measuring school quality, we would also need to build new *measures* of school quality. Once we reviewed the available data in Massachusetts, Rebecca and I discovered that data on most of what we needed to know was not currently being collected. In fact, we reviewed the available data across all fifty states and found that no state was gathering the full range of information we wanted. Even if we chose to draw on existing survey scales, we would want someone to thoughtfully guide us through that process. Hunter Gehlbach—a survey design expert then at Harvard University, who had recently accepted a position at Panorama Education—indicated an interest in helping.[12]

As the project took life, I wondered if Somerville was still the right place for it. The city obviously appealed to me because it's where I live and where my daughter goes to school. We also had strong civic support there—support that might have been difficult to generate elsewhere, given how untested our work was.[13] But perhaps we were making the convenient choice rather than the best choice.

Nevertheless, Somerville remained highly appealing as a research site. Perhaps the chief argument in its favor was its diversity. Somerville is one of the most diverse cities in the United States, with more than fifty languages spoken in the public schools. Forty percent of students are Hispanic, 35 percent white, 11 percent African American, and 9 percent Asian.[14] Half of the students in the city are from non-English-speaking homes. Sixty-seven percent are designated as low-income. Twenty-one percent have disabilities. These figures more or less

—-1
—0
—+1

match the demography of our nation's schools. Nationwide, roughly 50 percent of students are low-income and 12 percent are classified as Special Education students. Half are white, a quarter are Hispanic, 15 percent are African American, and 5 percent are Asian—figures that, according to projections, will grow to look more like Somerville over the next decade. According to the National Center for Education Statistics, by 2023, 45 percent of public school students will be white, 30 percent Hispanic, 15 percent African American, and 5 percent Asian.[15] In short, the city seemed like a highly reasonable test case.[16]

The second reason Somerville stood out was because of its size. Despite being the densest city in New England and one of the twenty most densely populated cities in the United States, Somerville is also quite small. There are only 75,000 residents, eleven schools, and 5,000 students in the city. That would make it possible to pilot our project across the entire district despite our being a small team. We would also be able to get major stakeholder involvement because the district's size would allow us to cultivate face-to-face relationships.[17]

Three years into this work, we have erected what I believe to be a formidable challenge to the test-based accountability system that dominates public education. Our model measures school quality far more holistically than any state system currently does, collecting roughly three dozen separate measurements, including a performance assessment of student knowledge and skills. We have developed a data visualization tool that offers more useful information and a more accurate picture of school quality than any existing state database. And, with support from the state legislature, we have assembled a group of districts—the Massachusetts Consortium for Innovative Education Assessment—that continues to push this work forward.

This book tells our story.

It is important to note, however, that this book is not merely a case study of one district. Rather, it uses my team's work in Somerville to

tell a larger story about how we think about school quality in the United States. All of us have ideas about what constitutes a good school. All of us have taken standardized tests. All of us care about school quality. And all of us have something to learn by thinking more deeply about how to assess the work of schools. Consequently, while the book draws substantially upon on-the-ground research in Somerville, it is designed for a national audience—of parents, educators, policymakers, and the public—interested in thinking through our current approach to educational data.

One way this book can be read is as background for better decision making. If we could value schools for the many things they do, we might reject the notion that there is a single "best school" out there to compete over. We might give up the pursuit of policies that diminish the mission of schools. We might stop ignoring a range of factors that are critical to school success. We might provide some relief to our most vulnerable schools—from the attacks that have magnified their burdens over the past two decades.

A second way the book can be read is as a field guide for improving the way we measure school quality. For those interested in more fairly and more comprehensively gauging educational effectiveness, this book will offer relevant information, instruments, and methodologies—a tool kit for parents, teachers, administrators, and policy leaders seeking to take concrete action.

We don't have all the answers. Our project, in many ways, is in its infancy, and each day we learn more about how to do this work the right way.

But this is also a matter of great urgency. We have two decades of evidence that current approaches to educational measurement are insufficient and irresponsible. Each day that we fail to act, we ignore the fact that we can do so much better.

—-1
—0
—+1

1

Wrong Answer: Standardized Tests and Their Limitations

IN 2014, 59 percent of students in Somerville public schools scored proficient or higher on the English Language Arts section of the Massachusetts Comprehensive Assessment System (MCAS).[1] Forty-eight percent were proficient or advanced on the math section.

Is that good or bad?

The numbers might be fine. But what about when we look at those numbers relative to the statewide average? Sixty-nine percent of students in Massachusetts scored proficient or higher on the English Language Arts section, and 60 percent scored at those levels on the Mathematics section.

Does that mean Somerville schools are academically underperforming? When Somerville's scores are compared with those for the state as a whole, that appears to be the case, so perhaps those numbers aren't as good as they seem at first glance.

What does proficiency measure, though? How was that cutoff point arrived at? Are students scoring below "proficient" unable to read and

compute? Are they inches—or miles—behind their higher-scoring peers?

And how can the influence of school quality be separated out? After all, factors such as family income and parental educational attainment are strong predictors of standardized test scores.[2] That being the case, how much responsibility does the school bear for the test scores of its students?

The answer to this last question can be incredibly difficult to sort out. A school populated by students with well-educated and affluent parents will produce much higher test scores even if it does a mediocre job of educating them. Should the school get credit for the high test scores of students who entered with higher levels of academic preparedness?

And wouldn't that logic apply to schools working with students who enter school with *lower* scores? Should schools be punished for working with students who arrive with lower levels of preparedness but who receive an excellent education? This is not to say that we should accept unequal outcomes for students from less privileged backgrounds. It does, however, raise questions about what is realistic. It also raises questions about how much we can rely on test scores to gauge the quality of a school's educational program.

Two-thirds of students in Massachusetts are white; one-third of them are members of racial minorities. In Somerville, the reverse is true. Statewide, 18 percent of students speak a language other than English as their first language. In Somerville, that figure is nearly three times larger: 50 percent. And the rate of economic disadvantage in Somerville—36 percent—is ten points higher than it is in the rest of the state.

All of this means that, statistically speaking, we should expect to see much lower test scores in Somerville. Again, we should not establish lower expectations for low-income students, students of color, and nonnative speakers of English. But we should probably recognize

—-1
—0
—+1

that even if their scores are lower than those of their more privileged peers, their schools may be doing a good job. Blaming the schools, particularly when they may be the only social institution holding up their end of the bargain for these young people, seems counterproductive.

If we account for the student body, then, perhaps Somerville schools are doing incredibly well. Students who begin their academic careers at a disadvantage make up significant ground, at least insofar as academic ability is measured by those tests. Those who enter school with no significant disadvantages appear to do as well as students anywhere.

Of course, even if we look at the scores of various subgroups—across race, income, language, gender—we are still only looking at test scores. And what do those actually measure?

What the Tests Measure

Critics of testing sometimes make the case that standardized tests are not accurate measures of student learning. Although they aren't entirely wrong, this is an overstatement of the truth. Generally speaking, if a student scores below "proficient" on a measure of grade-level math or reading, that student is not thriving academically. This may not be the school's fault—the student may have arrived several grade levels behind his or her peers. Nevertheless, it is useful information for educators and policymakers as they seek to track student progress and allocate resources.[3]

Still, standardized tests are quite limited in what they can tell us about student ability. Consider a story from the educational psychologist Lee Shulman, recalling his own education:

> When I was an undergraduate at the University of Chicago in the late 1950s, I attempted to cram for the end-of-year comprehensive

> examination in the history of western civilization—a nine-hour multiple-choice and essay test. I thought I had done quite well on the exam and was thus shocked to receive a "C" for the course. I asked to meet with a member of the Evaluation Office to learn why I had performed so poorly. We sat down and examined my performance, using Bloom's taxonomy as a template. I had "aced" the multiple-choice section, with its emphasis on recall; cramming can be a pretty good strategy for remembering facts and ideas, at least over the short term. But I had simply not studied well enough to integrate the ideas and to be able to synthesize new interpretations and arguments using the knowledge I had crammed into my head.[4]

If the evaluation of Lee Shulman's knowledge had been limited to what could be measured by a multiple-choice test, he would have appeared to be a highly accomplished student of history. What was revealed instead, however, was far more complex. He may have known the facts of history, but he hadn't learned to use them in the service of new ideas or to solve novel problems. Only something far more complex than a multiple-choice test—a series of essays, graded by human beings—could determine that.

Machine-scored multiple-choice questions are the main tool we use to measure learning in K–12 public schools. Why? Because they are cost-effective and easy to standardize. But they place far too much emphasis on memorization, and far too little on complex cognitive processes like problem solving. In fact, they tell us very little about what is going on inside students' minds.[5]

Consider a sample reading comprehension passage from the fourth-grade California English Language Arts test:

> Long ago, when the world was new, Beaver had a long, thin tail. He loved to dive, but his long tail didn't help him get to the bottom of the pond fast enough. He couldn't use his tail to slap the mud into place when he built a dam. One day, Muskrat swam by. Beaver

—-1
—0
—+1

> noticed Muskrat's broad, flat tail. He realized it would be perfect for diving and building dams. At the same time, Muskrat gazed enviously at Beaver's tail. Muskrat loved to swim fast, and his broad, flat tail dragged in the water and slowed him down. He thought it would be better to have Beaver's tail. So Muskrat said, "Beaver, I would do anything to have a tail like yours." "Is that so?" replied Beaver. "I was just admiring your tail. Why don't we trade?" Muskrat eagerly agreed, and they exchanged tails right then and there.

On one hand, if a student cannot answer basic questions about this passage, he or she is likely not a strong reader. On the other, a student who answers questions correctly does not necessarily possess all of the skills and habits of mind we might wish to cultivate. What are the characteristics of a strong reader? Speed? The ability to pick up on tone and nuance? Attention to authorial style and word choice? None of this is measured by the test.[6] Instead, the answer bank presents four options:

> What is the main event in "The Tail Trade"?
>
> a) Beaver builds a dam with his tail.
>
> b) Muskrat and Beaver exchange tails.
>
> c) Muskrat and Beaver try out their new tails.
>
> d) Beaver slaps his tail to warn of danger.

A student who cannot answer "b" is struggling, but beyond that, the question doesn't tell us much about the things that matter most—fluency, interpretive skills, analytical ability, or precision. Nor can the question help us to diagnose what a student needs in terms of instruction.

It is also worth noting here that even a question as seemingly straightforward as this one is not necessarily fair for all students. A student whose first language is not English, or who for some other reason has a limited vocabulary, may struggle to distinguish between

-1—
0—
+1—

"b" and "c" despite being a perfectly good reader. If he or she is unfamiliar with the word "exchange," that student will face a coin flip situation and have to guess.

Test questions, of course, can be much worse than those in this example. Several years ago in Massachusetts, a question on the state exam prompted elementary-level students to write a fictional story about a mysterious "trunk" they had stumbled upon. Naturally, as former Somerville superintendent Tony Pierantozzi recalled, "we had English language learners writing about elephants, cars, and swimsuits." The state's intention was for students to describe what would be in a box.

"That wasn't as bad as a question I remember from my own schooling experiences," Pierantozzi added. "It was an analogy: Caesar is to salad as yacht is to . . . what? The answer was boat. But I had no idea what a yacht was." Tony's experience was not exceptional. Research by scholars such as E. D. Hirsh makes a powerful case for the relationship between content knowledge and reading comprehension, and when it comes to reading passages on standardized tests, the inclusion of particular content can skew results dramatically. Largely, it is skewed in favor of white native English speakers from middle- and upper-income households.

Standardized tests favor culturally dominant students in other ways as well. The tests, for instance, are written in standard English, which is often not the English encountered by low-income and minority students at home. As a result, while some students may know how to read and write as well as their more privileged peers, they may be less familiar with particular phrasing. Or they may select phrasing that, though perfectly acceptable in their neighborhoods, reads as "wrong" on a test. Relatedly, test writers often assume that students are familiar with objects such as saucers or hampers that, though common in many homes, are not present in all of them, or that go by different

—-1
—0
—+1

names in different places. Or, perhaps even more unfortunately a student may respond with a right answer but still be marked wrong because it is not the *best* answer—the one the test publisher is looking for.[7]

Much of this, of course, is accidental—a product of test writers' crafting a product with a generic audience in mind. But some practices are glaringly problematic. When test items are measured for their psychometric properties—how well they measure what they are trying to assess—the items on which lower-scoring groups end up scoring higher are usually dropped from the test. In other words, if the kids who score worse—generally low-income kids and students of color—actually score *better* on a question, that question often gets cut from the test. The reason for this is that such outcomes do not align with results from the test as a whole. Consequently, those questions are viewed as aberrations in need of "smoothing." It is also worth mentioning here that many questions are never included in tests because too many students would answer them correctly. The test, after all, wouldn't be useful if everyone aced it. Yet this approach to design certainly distorts the picture of student achievement—sending a message that students know less than they actually do.[8]

Some problems with standardized tests don't have anything to do with the kinds of questions that get asked on them. For instance, we conduct examinations in a manner that requires students to sit for a prolonged period of time in silence—a format that squares more with the backgrounds and experiences of some students than others and has little in common with the world of work for which they are being prepared. While there may be some benefit to the skill of sitting quietly to focus on a test, that skill is separate from one's ability to read, write, and compute. Nevertheless, the ability to sit still does translate to higher test scores.[9]

Relatedly, cultural conceptions of young people, shaped by factors like race and class, can also affect test scores. A substantial body of research, for instance, indicates that students from minority backgrounds tend to experience "stereotype threat" when sitting for exams—fulfilling the high or low expectations that have been set for them. African American students, for instance, score lower when they feel that a test is a measure of their intelligence and when they worry that their performance will be viewed through the lens of racial stereotypes. Conversely, students who are stereotypically depicted as high achievers—particularly Asian American students—score higher in such situations. As surprising as it might sound, this phenomenon has been confirmed repeatedly by experts.[10]

Perhaps most significantly, there is the fact that tests often tell us more about student home lives than about schools. The two top predictors of student standardized test scores tend to be parental education and family income. Why? Not because higher-income families headed by college graduates can buy more books, which they certainly can, or afford tutors, though they can do that as well. Instead, the most significant impact of those factors seems to be an indirect one. From birth, families pass on to children a set of values and beliefs about the importance of school—signaling what is desirable and setting expectations about academic achievement. Parents with high levels of educational attainment and success communicate a particular worldview to their children, setting standards not only through their words but also through their actions.[11]

Especially before the start of formal schooling, families are also the primary teachers for children. In those early years, much of a child's cognitive development depends on the interactions he or she has with family members. Families from more privileged backgrounds, and who have resources at their disposal, tend to engage in more verbal interaction with children, provide more cognitive stimulation, promote

—-1
—0
—+1

particular kinds of learning strategies, and establish stimulating environments. Love, of course, matters a great deal, and no particular demographic group loves children more than any other. But when it comes to educational achievement, privilege matters tremendously. For quantitativeevidence of this fact, one need only count the words children hear in different settings. As one longitudinal study found, by age three, children from more privileged families hear thirty million more words than their less privileged peers.[12] No wonder that when it comes time to take standardized tests their scores are so much higher.

All of this makes testing problematic across racial, ethnic, linguistic, and cultural dimensions As a result, we have many reasons to question how much standardized test scores are telling us—especially when we recognize what the tests are *not* measuring.[13]

Take the case of Jones Elementary in Arkansas, which earned a D rating from the state for its student standardized test scores. This reflects the population of the school, where 98 percent of students are from low-income households and 80 percent are English language learners. Yet the school has been recognized by the U.S. Department of Education for academic growth, faculty collaboration, its school-based community health clinic, and, somewhat ironically, its effective use of data. The school has also initiated a home library program and a "parent university" program to help offset the effects of demography. In short, it is doing a tremendous amount to support the young people in its care.[14]

How, then, can we even pretend to determine a school's quality by looking only at its test scores?

Imagine that we have done a kind of calculus and concluded that, given the school's population, it is doing well according to the tests. Imagine that we have figured out what those tests tell us about the basic academic competencies of students, and imagine that we main-

tain a healthy skepticism about all of it. This would be quite an accomplishment. We would nevertheless have to ask: How complete is our picture of the school?

Remember: the tests used to measure school quality tend to assess only the domains of English Language Arts (or, as previous generations knew it, English) and Mathematics. Thus, while we have performance data for student work in those two subjects—limited data, at that—we lack such data for other areas. In addition, those tests tend to be given only in grades three through eight and once in high school—a highly incomplete picture, indeed.

One option, if we wanted to solve this problem, would be to expand testing.[15] Examining students in all subject areas, and at all grade levels, though, would take an inordinate amount of time—doubling or even tripling current commitments to testing. Imagine how exciting this prospect would be for the average student, who already may think school is a pencil-dulling drag. More disturbingly, consider the opportunity cost of testing: while testing is happening, no new learning can take place.

Additionally, there is the problem that not all school subjects are conducive to standardized testing. It is foolish to test a student's ability to think like a historian by asking multiple-choice questions. Historians, after all, don't sit around reciting facts; instead, they try to solve puzzles by weaving together fragments of evidence. For subjects such as art and music, the prospect of gauging ability through multiple-choice tests is even more absurd.

Yet even if we possessed student test score data for all subject areas, we still wouldn't have a complete picture of the schools.

We would have no data about how happy students are or how challenged they feel.

We would have no data indicating how safe they are in the hallways or how cared for they are by the adults in the school.

—-1
—0
—+1

We would have no data on how creative they are or how hard they're working.

We would have no data on how healthy they are—physically, socially, or emotionally.

We would have no data on how many of them can play a musical instrument or program a computer or design an experiment.

In short, we would have no real picture of what life inside a particular school is like, and isn't that what we really care about?

The harshest critics are wrong that standardized tests are worthless. Tests tell us something basic about student achievement, and that information can be critical in advocating for additional resources and supports.

But critics of testing are right about a great deal else. Standardized tests are limited in what they reveal and are not entirely fair across racial, ethnic, and economic lines. They take up too much instructional time and overlook most of what we value in schools. Relying on test scores as a measure of school quality, it seems, is a fool's errand.

How Did We Get Here? A Brief History of Testing

If standardized tests are so limited, why do we use them?

As with many conventions in education—grouping children by age, closing school during the summer, assigning A–F letter grades for student work—history plays a significant role. We accept these things because they are standard, routine, familiar. Such features are so conventional as to go unremarked upon—noticed more when they are absent than when they are present.[16]

Standardized testing is cheaper than many alternatives. But we accept it because it has long been the norm. Sure, it may be odd to gauge student knowledge through a series of test-bubbles darkened by

number two pencils. But testing is a fundamental part of education in the United States. It has been for generations.

All traditions have origins, though, and understanding those origins can help us cultivate a critical perspective toward what we might otherwise simply accept. This is no less true of standardized testing.[17]

At the end of the nineteenth century, educational policymakers sought to build stronger systems of governance as well as to make education more scientific—mimicking their colleagues in medicine. Having succeeded in creating a system of public schools during the previous generation, they were anxious to establish mechanisms that would give them more direct governance over those schools. In other words, it was not enough to have created statewide networks of schools that would be free and open to all children. They also wanted to control what was going on inside those schools.

State and municipal policy leaders in the period after the Civil War were interested in asserting control particularly over school principals, who had long operated with very high levels of autonomy and who tended to possess more power than administrators higher up in the system. To wrest this power from principals, experts and policymakers needed some mechanism for setting educational goals, tracking school progress, and casting judgment. Yet the bulk of their information about school progress came from these same principals. Standardized tests, then, promised to shift the balance of power—giving policy leaders a yardstick for measuring school performance and making centralized governance feasible. The practice quickly spread.[18]

The first standardized tests included short essays in the mix of questions. Over time, however, writing prompts were replaced by straightforward questions of fact. After all, it is far easier to measure whether a student knows who wrote the Gettysburg Address than to

—-1
—0
—+1

measure his or her ability to think critically about the speech. It is even easier to score questions that eliminate student writing entirely, and multiple-choice questions, developed in the early twentieth century, did just that. By asking students simply to circle correct answers, they eliminated the possibility of human subjectivity in scoring.

A great deal of excitement surrounded such advances in standardized tests. Even in the early years of testing, though, there were critics. A math department leader complained in 1927 that it was "quite possible to drill for an examination and to pass a large number of pupils with high ratings without giving any breadth of outlook or grasp of underlying principles."[19] And Henry Linville, the president of the New York teachers' union, in 1930 observed that the state's overemphasis on testing "indicates a frightful standardization of curriculum, of methods and of objectives."[20]

Perhaps the strongest pushback against standardized tests, at least prior to the twenty-first century, came in the late 1930s from the Educational Policies Commission (EPC). Formed by the National Education Association and the American Association of School Administrators, the EPC sought to pinpoint the features of American schools worth fighting for—in the wake of economic depression, and on the edge of another great war. In 1938 the EPC released *The Purposes of Education in American Democracy,* in which the group concluded that "most of the standardized testing instruments" used in schools had failed to address "the development of attitudes, interests, ideals, and habits." Instead, the tests focused "exclusively on the acquisition and retention of information." In the broader picture, that seemed to them "relatively unimportant."[21]

What the members of the commission wanted to see instead was a broader set of measurements that aligned with the facets of the American educational system valued by a democratic people. As they wrote:

> Measuring the results of education must be increasingly concerned with such questions as these: Are the children growing in their ability to work together for a common end? Do they show greater skill in collecting and weighing evidence? Are they learning to be fair and tolerant in situations where conflicts arise? Are they sympathetic in the presence of suffering and indignant in the presence of injustice? Do they show greater concern about questions of civic, social, and economic importance? Are they using their spending money wisely? Are they becoming more skillful in doing some useful type of work? Are they more honest, more reliable, more temperate, more humane? Are they finding happiness in their present family life? Are they living in accordance with the rules of health? Are they acquiring skills in using all of the fundamental tools of learning? Are they curious about the natural world around them? Do they appreciate, each to the fullest degree possible, their rich inheritance in art, literature, and music? Do they balk at being led around by their prejudices?[22]

Without a doubt, these are all critical questions. But no tools existed to easily, inexpensively, and systematically address them. One can imagine policymakers nodding in agreement and then asking, frankly, "How on earth would you go about measuring these?" Projects like the Progressive Education Association's "Eight-Year Study" were designed to bring a degree of balance to assessment in American classrooms.[23] Ultimately, however, the ease and efficiency of standardized tests were too much to overcome, particularly in the late 1930s, when the first automatic test-scoring machines—including the one patented by IBM in 1937—became available.

Tests also served a variety of other purposes that made them seem indispensable. Perhaps chief among those purposes was structuring educational meritocracy.[24] Of course, American schooling has never

—-1
—0
—+1

been truly meritocratic; one need only examine the strong correlation between income and achievement to understand this fact. But unlike in other nations, where educational opportunity was limited to those of means, the American educational system was founded on the premise that it would serve as a "great equalizer." Open to most students and funded by tax dollars rather than by tuition, the schools would provide upward mobility to the poor and, in so doing, benefit those individuals and society as a whole.[25] Yet such a vision required a mechanism for reaching across schools and districts to identify top achievers. As one 1916 observer wrote, "Fairness in the award of honors, justice in determining failures and dismissals, and incitement of the student to better work can be attained only to the extent to which a common standard for the awarding of marks is understood, accepted, and acted upon."[26] How could such a common standard exist, though, without a common test?

Desperate for a fair way to identify the best and brightest students, many policymakers saw standardized testing as an important technology: a clear and seemingly unbiased way of comparing students against each other. As the author of a 1922 textbook wrote, "When such scores are represented by a simple graph, say with one line showing the given pupil's attainment in these tests and another line the attainment of the average American child of this age or grade who has taken these tests, then the pupil has his strong and weak points set before him in a manner that is perfectly definite and objective."[27] In short, standardized tests seemed to be an essential ingredient in creating a fair playing field. Of course, things wouldn't turn out that way. But there was a kind of hope, at least on the part of many, that testing would help identify and nurture raw talent.

As alluded to earlier, standardized tests also made governance easier. For decades, policy leaders had maintained little direct control over the schools, which operated with a significant level of autonomy

with regard to curriculum, instruction, and assessment of learning. Consequently, it was impossible for policy elites to compare the work of teachers or students across schools. Their powers were largely limited to issuing recommendations.

Standardized tests changed this state of affairs by allowing policymakers and legislators to measure institutions from afar. Shortly after the development of the New York State Regents exam, for instance, policymakers began using the test to expand their direct control over the schools—primarily by tying funding to performance. "Successful" schools were rewarded with a funding carrot; their less successful counterparts received the stick. Still, not everyone took the Regents exams. As an investigator for the New York Board of Estimate and Apportionment put it in 1912: "The whole problem is one of lack of measurements and lack of definite tests . . . And as a result we go on year after year without looking for any positive checks on the results we obtain."[28] Even half a century later that would remain an issue. As a 1971*New York Times* story reported, only by expanded testing could the state provide "an objective, uniform standard of attainment . . . from one locality to the next."[29] Only by conjuring a way to efficiently and uniformly "see" school performance—in this case, through test scores—could officials in central offices and state capitals establish enough systemic order to make governance possible.[30] Only then could they wield any real control.

Not surprisingly, the scores produced by such tests became increasingly important to policy and governance structures over time. As a 1954 *New York Times* story reported, the Regents tests, according to the state education department, were by that time being used "to measure pupil achievement, to serve as a basis for admission to college, and . . . [to] improv[e] the quality of instruction in the major secondary school subjects."[31] Because policy leaders were reliant on test scores to pursue those critical aims, they may have concluded

that they simply couldn't do their jobs without standardized tests. Nor could they exert the kind of authority they had become accustomed to.

By midcentury, standardized test scores were a common currency in educational policy—the information underlying efforts to measure and control schools. This was disturbing to some, of course. As one critical scholar wrote in 1959: "It is regrettable that many schools appear to appraise teaching competence by making comparisons of groups of pupils on the results of standardized achievement tests. When teachers realize that their teaching effectiveness is being evaluated by this method, many of them find ways of teaching for the tests, thus reducing possible contributions the tests might make toward genuine program improvement."[32] Whatever their flaws, though, test scores represented the foundation of an increasing number of educational policy structures. And they remained an incredibly simple and straightforward tool for calculating school quality. As a 1970 report by the New York State Education Department put it, "The distributions of scores are processed by computer, and reports which summarize the results in conveniently interpretable form are returned to each school and central office."[33] Never mind their weaknesses. As a form of currency, they worked. They were easy to compile and easy to compare.

Standardized tests also won a foothold in American education because an entire corps of professionals arose around the standardized testing apparatus. The first wave of testing experts entered K–12 schools at the end of World War I. In the army, they had worked to develop and administer military intelligence tests used to sort and rank recruits—tests that scholars would later reveal to be highly problematic.[34] At the time, however, these military-trained testing specialists were presumed to possess a sophisticated set of skills, and they were returning to civilian life in search of work. In the eyes of many state and district leaders, the timing couldn't have been better. As one

observer noted, "The fact that two or three hundred young men who have for several months been working in the psychology division of the Army are now about to be discharged offers an unusual opportunity for city schools to obtain the services of competent men as directors of departments of psychology and efficiency, for such purposes as measuring the results of teaching and establishing standards to be attained in the several school studies."[35] Soon every state office of education, as well as every large district, had its own staff of testing experts.

As the work of these experts grew more complex and specialized, they played an increasingly central role in testing. Given their positions, they were unlikely to focus on the fundamental limitations of standardized testing. Instead, they were far more likely to tout the incremental improvements being made—to the tests themselves, as well as to increasingly sophisticated analysis techniques for interpreting scores. Their science, as they understood it, was both highly precise and rapidly evolving.

Standardized tests were also highly profitable. And that meant that however many critics lamented the existence of such tests, entrepreneurs would work not just to meet demand but also to *manufacture* it—convincing schools and districts that testing was essential to progress. Already by 1920, plenty of individuals and organizations were making their livelihoods from the production and scoring of tests. They were scrambling to capture as much of the market for testing as they could in those early days.

Selected Tests from 1920 **(listed alphabetically)**

Baldwin's Public School Music Test
Barnard's Test in Roman History
Barr's Diagnostic Tests in United States History
Bell's First-Year Chemistry Test

—-1
—0
—+1

Caldwell's Science Tests
Chapman's Physics Test in Electricity and Magnetism
Clemens's Grammar Test
Coleman's Scale for Testing Ability in Algebra
Courtis's Dictation Spelling Tests
Grier's Range of Information Test in Biology
Handschin's Foreign Language Tests
Harlan's Test of Information in American History
Henmon's French Tests
Hillegas's Scale for the Measurement of Quality in English Composition for Young People
Lohr's Latin Test
Manuel's Series of Tests for Studying Talent in Drawing
Minnick's Geometry Tests
Rogers's Mathematics Tests
Rugg's Tests for Historical Judgment
Starch's Tests for Measuring Grammatical Knowledge
Thorndike's Algebra Test

All told, roughly 250 tests were commercially available in the early 1920s, and that number continued to grow.[36] Providers multiplied until testing became truly big business—a business so lucrative that large organizations began to consume their smaller competitors.

By the 1950s testing had become a major industry—dominated by a handful of increasingly influential players. According to the National Board on Educational Testing and Public Policy, test sales to K–12 schools in 1955 were $7 million. Fewer than ten years later, that figure rose to $25 million.[37]

As the educational system grew larger, as students stayed in school longer, and as more schools and districts began tracking student performance via standardized tests, the market continued to expand. By

the end of the twentieth century, standardized testing of elementary and high school students—to say nothing of college entrance examinations and the like—was a $300 million industry.

Industry growth was hardly at its apex. In 2013, Pearson, a British publishing company with wide reach in the U.S. market, posted revenues of roughly $10 billion, half of which came from the North American market, and much of which came from state and federal testing contracts. The company inked a five-year testing contract with New York worth more than $30 million. It won testing rights in Texas—a deal worth nearly half a billion dollars—and it secured contracts with roughly half the states for statewide testing. Pearson gets this business because it's big—it has purchased many smaller testing and publishing companies. But it also secures these contracts because it wields significant influence in Washington and in state capitals. Between 2004 and 2014, for instance, Pearson spent roughly $7 million lobbying at the federal level, and far more at the state level.[38]

Pearson is not alone. CTB / McGraw-Hill, a Pearson competitor, also has revenues in the billions, with major contracts in roughly two dozen states. And like Pearson, CTB / McGraw-Hill has also learned to wield its influence in the halls of government.

What about the public? Policy leaders, testing experts, and the testing industry had their reasons for continuing to support the use of standardized tests, but why did the public never demand an end to the testing regime?

One factor that kept opposition to testing relatively muted is that those best positioned to publicly decry the tests—leaders in government, business, and school systems—were often uncritical of them. This should not be surprising, as success in school frequently serves as a sorting mechanism for placing individuals in positions of influence. Naturally, as a result, many see their places at the top of the

—-1
—0
—+1

socioeconomic hierarchy as a product of their intelligence and hard work. After all, it is much harder to understand the inadequacy of tests if one has always succeeded on them. Rare are cases like that of the prominent American psychologist Robert J. Sternberg, who struggled with tests and still managed to achieve a position of influence—a position he then used to debunk our widely accepted notions about measuring intelligence.

Another key factor is time. Over decades of experience, the American public became accustomed to tests, to the numbers produced by tests, and to the narratives that arise around test results. By the 1920s, tests were already a routine part of life in school for teachers and students. As the *American School Board Journal* reported in 1922, "Measurements of achievement, through the use of educational tests, have come to be a common feature of the public schools."[39] Over time this situation only became more prevalent. A college-bound student in 1960, for instance, might take district- and state-mandated tests, diploma tests, Advanced Placement tests, and the SAT before arriving in college. The non-college-bound student might win a reprieve from a few of those, but after graduation from high school, there would likely be more standardized tests to face in the world of work: tests to become a licensed beautician, nurse, auto mechanic, or law clerk. Tests could not be avoided.

Ask almost anyone on the street, and they will acknowledge the importance of tests. They may not like them and may not be in favor of an overabundance of tests, but they accept them. As one Somerville community member observed: "I just don't want my kid getting tested all day instead of learning. But I do want him to do well on tests. I do want him to succeed." And as Somerville congressman Michael Capuano put it in a speech that was otherwise very critical of standardized testing: "Tests are just a part of life. You can't avoid tests."[40]

In short, constant exposure to tests over the years has led to general comfort with and acceptance of them. Culture is always evolving. And as each successive generation of Americans has come of age, they have done so in a world in which testing was increasingly natural—a part of normal life. One could certainly *dislike* testing, just as one can dislike sitting in traffic. But to imagine a world without testing is a different matter entirely.

Of course, it is also true that for most of the twentieth century, tests weren't as invasive as they are today. The standards and accountability movement, which culminated in the 2002 No Child Left Behind (NCLB) law, would change that.

A Nation at Risk and the Dawn of Test-Based Accountability

In the 1970s, schools using standardized tests to measure student achievement usually opted for exams like the Iowa Test of Basic Skills or the Stanford Achievement Test. Developed for a generic national audience, such tests were unrelated to any particular school curricula. Instead, they measured student knowledge of what was presumed to be basic or common content. Rather than weigh in on how many questions students should be able to answer correctly, test developers generally measured student performance relative to the scores of other students. Consequently, such tests tended to be used for diagnostic purposes rather than for accountability.

That would begin to change in the last decades of the twentieth century. In the 1970s, frustration with a perceived decline in educational quality led policymakers to seek a means of holding students accountable for their learning—via minimum competency testing and high school exit exams. Between 1972 and 1985, the number of state-level testing programs rose from one to thirty-four.[41]

—-1
—0
—+1

Policy elites had long sought to pry open the classroom door to gain greater control over teaching. For generations they had been writing new curricula, setting new goals, providing more professional training, and pouring new resources into schools—all without much impact on the daily work of teachers, who could simply close their doors and ignore any reform efforts they deemed unsuitable for their classrooms. Sometimes this was prudent; teachers kept out ill-advised policy efforts. Sometimes, no doubt, this inhibited growth. But whatever the ultimate impact, policy elites were tremendously frustrated by the level of constancy in teacher practice across time.[42]

State-run testing offered a way of solving the problem of teacher autonomy. If the state were to create clear student learning standards and develop tests aligned with those standards, policy leaders might gain the power that had eluded them. They still might not be able to control what teachers were doing inside their classrooms, but if they could measure the degree to which the state-designed curriculum was being learned, they could light a fire under the feet of teachers and school administrators. This theory soon took on the moniker "standards-based accountability."[43]

Insofar as standards-based accountability would give policymakers new powers, it had a strong appeal. But it had a particular allure during the late 1970s and early 1980s, as rhetoric about a crisis in American public education escalated. Concerns about a deteriorating economy, the ongoing Cold War, the rise of foreign competitors like Japan, and declining SAT scores led to the assertion that schools were in trouble and that America's position in the world was in jeopardy. Perhaps most famously, *A Nation at Risk,* a report issued in 1983 by the National Commission on Excellence in Education, warned:

> Our once unchallenged preeminence in commerce, industry, science, and technological innovation is being overtaken by competitors

-1—
0—
+1—

> throughout the world . . . We report to the American people that while we can take justifiable pride in what our schools and colleges have historically accomplished and contributed to the United States and the well-being of its people, the educational foundations of our society are presently being eroded by a rising tide of mediocrity that threatens our very future as a Nation and a people. What was unimaginable a generation ago has begun to occur—others are matching and surpassing our educational attainments.[44]

A slew of similar reports followed. Typical was that of the Committee for Economic Development—a group of 200 business executives and educators—which claimed in 1985 that "Japanese students study more and learn more. They spend more time in class than their American counterparts do; and by the time they graduate from high school, they have completed the equivalent of the second year at a good American college. In science and mathematics, Japanese test scores lead the world."[45] Standards-based accountability would give policymakers the tools to ratchet up their demands on schools to address this alleged crisis.

The first step toward a true standards and accountability movement was articulated by Tennessee governor Lamar Alexander in 1985. As he put it: "The Governors want to help establish clear goals and better report cards, ways to measure what students know and can do. Then, we're ready to give up a lot of state regulatory control—even to fight for changes in the law to make that happen—*if* schools and school districts will be accountable for the results."[46] Three years later, Republican presidential candidate George H. W. Bush made this idea the centerpiece of his educational policy agenda. Noting that he wanted to be "the Education President," Bush promised to lead a "renaissance of quality" by working with the nation's governors to devise more rigorous educational standards.[47]

—-1
—0
—+1

Less than a year after his election, Bush and the National Governors Association co-organized the 1989 Charlottesville Education Summit. At the heart of the summit was the work done by a committee of governors, led by Arkansas governor Bill Clinton, who were seeking to exert more influence over education. In just two days, the committee hammered out a basic framework for educational standards.

Over the next several months, the Bush administration, along with Clinton and other members of the National Governors Association, crafted a piece of legislation—America 2000—that Bush proposed in his 1990 State of the Union address. The plan called for voluntary national standards and tests. Yet Congress, which had not participated in the Charlottesville summit, sank the bill. Not long afterward, Bush lost his bid for reelection.

Fortunately for backers of standards-based accountability, the candidate who defeated Bush was Bill Clinton, who swiftly resurrected America 2000, rechristened it Goals 2000, and shepherded it into law in 1994. Though Goals 2000 did not mandate testing or authorize consequences for low performance, it did establish a federal interest in standards and accountability. The goal for student achievement and citizenship, for instance, stated, "By the year 2000, all students will leave grades four, eight, and twelve having demonstrated competency over challenging subject matter including English, mathematics, science, foreign languages, civics and government, economics, arts, history, and geography, and every school in America will ensure that all students learn to use their minds well, so they may be prepared for responsible citizenship, further learning, and productive employment in our Nation's modern economy." The new law further specified goals for increases in academic performance and reductions in achievement gaps.[48]

Perhaps more importantly, the law also provided grants to help states develop content standards. By the time Clinton left office in Jan-

uary 2001, most states had established academic standards, and many had begun to assess students through statewide testing. A clear foundation had been laid for the architects of No Child Left Behind.

NCLB, ESSA, and the Era of High-Stakes Testing

No Child Left Behind was the first piece of legislation pursued by George W. Bush upon assuming the presidency in 2001. Continuing work initiated by his father a dozen years earlier, Bush also modeled the new law after his own work in Texas, where as governor he had strengthened the state's school accountability system. As a 1999 *New York Times* story put it: "The key to the Texas accountability system since it was instituted in 1991 has been relentless focus on testing. Every year, from third through eighth grade and once again in high school, virtually every Texas public school pupil takes a version of the Texas Assessment of Academic Skills . . . The results, along with school attendance figures, determine a school's rating in the state."[49]

NCLB had bipartisan appeal in that it advanced projects of previous Republican and Democratic administrations. And evidence from Texas seemed to support the model of standards-based accountability. Sure, there were those who complained about the focus on testing in Texas. As one mother observed: "In many years, all my daughters' teachers have done is drill them for [the test] instead of giving creative writing or interesting projects . . . The system may look good on paper, but I feel my daughters are getting ripped off."[50] Yet student achievement scores in Texas had gone up steadily under Bush's governorship. Of particular interest to lawmakers were the impressive scores of traditionally underrepresented minorities. Flanked by Democratic senator Ted Kennedy, Bush signed NCLB into law on January 8, 2002.

NCLB was the culmination of a federal legislative process that began a dozen years earlier in Charlottesville, Virginia, building on

both America 2000 and Goals 2000. But NCLB was also erected on an even older foundation. Though technically a new law, NCLB was actually an updated version of a law passed several decades earlier—the Elementary and Secondary Education Act (ESEA). Signed by President Lyndon Johnson in 1965, the ESEA was conceived of as a part of the War on Poverty—a way of channeling federal funds to schools with relatively high rates of low-income students. The primary component of the law, Title I—"Financial Assistance to Local Educational Agencies for the Education of Children of Low-Income Families"—outlined a model in which federal funds would be distributed to state departments of education. States would then allocate resources to districts, and districts would provide funds to schools. Today, Title I funds reach roughly half of the nation's schools.

At the time of the original law's creation, when school funding formulas were even more inequitable than they are presently, these funds were particularly important. Even today, Title I funds are a critical source of revenue for schools with high percentages of students living in poverty. American schools rely heavily on local property taxes for their funding—nationally, 45 percent of funds come from local sources—meaning that high-poverty neighborhoods often struggle to adequately fund their schools. And though state funding, which on average accounts for another 45 percent of spending, can be structured to address the inequities in local funding, most states fail to achieve budget parity. Consequently, though the federal contribution is fairly small—accounting for roughly 10 percent of school budgets, on average—it is not insignificant. Some 50,000 K–12 schools count on those federal dollars to fund daily operations.[51]

The federal government has no direct constitutional authority over the schools. Instead, that power belongs to the states. In accepting federal funds, however, states agree to particular terms with regard to how that money will be used. In previous iterations of the ESEA, the

federal government asked little in return for financial support. Schools could count on receiving federal dollars to alleviate the impact of poverty, and were responsible largely for showing that the money had been spent appropriately.

NCLB, however, would change that.

The key component of NCLB was test-based accountability. A reality for states like Texas prior to the law's passage, the new mandate was a shock for others. Specifically, the reauthorized law required states to conduct annual standards-based testing in math and English for all students in grades three through eight and one year in high school. States would control their own standards documents and define their own levels of "proficiency"—producing great disparity across states in defining academic competence—but all states would be required to bring 100 percent of students to a level of proficiency within twelve years of the law's passage.

Soon, all fifty states began testing students in math and English at a minimum of seven different grade levels. And those test scores carried significant consequences. Schools that did not meet targets—"Adequate Yearly Progress" in the language of NCLB—were to be sanctioned in accordance with the law's guidelines. A school that failed to meet targets for more than two years in a row, for instance, would be required to notify parents of its failures. A school failing to meet targets for over five years would be subject to closure. Never had standardized tests been so high-stakes for so many.

NCLB was the culmination of two decades of work. First, policymakers and the public had to be convinced that it was important to produce a quantitative picture of how the schools were doing. Then the states had to develop standards. Then states needed to develop tests aligned with those standards. Finally, once all of these pieces were in place, the muscle could be introduced—accountability for results. Still, as lawmakers would learn, it was far from a finished product.

—-1
—0
—+1

The massive scale of testing, coupled with NCLB's reporting requirements, produced mountains of data—in many cases issued through school-level "report cards." Note, though, that those report cards were quite limited. As one nonpartisan group put it, "The metrics, weights, formula and report card do not reflect public values."[52] And, as one might expect, many of the data, from the outset, did not look good. What surprised many lawmakers, though, was that scores did not seem to be improving much over time. The theory of action behind standards-based accountability, after all, was that educators would work harder to achieve the results expected of them, and that parents would exert more pressure on the system to improve. Whatever the initial levels of performance, the theory predicted widespread change.

Yet NCLB came with only a small increase in funding to improve outcomes, particularly given the new strings being tied to the federal revenue stream. It provided only weak supports for schools, such as a mandate to hire largely ineffective tutoring companies. And it came with little guidance about how to interpret data for parents.[53]

There were other obvious limitations to the new law. States could insist upon higher test scores, for instance, but they could not control any other aspect of school life, for which no data were available. In fact, they generally lacked data on anything other than math and English scores at seven grade levels. What was happening in the history classroom or the science classroom? What was going on in second grade? Or eleventh?

States also had no means for controlling the manner by which schools sought to raise test scores. It was obvious that some schools were narrowing the curriculum or emphasizing test preparation over other forms of instruction. Yet how could they be told not to when those were the only factors for which they were being held accountable? Many policymakers chose to put on blinders and assume that

schools with high test scores were doing a good job. Others, however, worried that the new law was actually encouraging counterproductive practices.[54]

Some policy leaders at the state level also recognized early on that NCLB's accountability mechanisms offered little in the way of diagnosis. Low test scores on reading comprehension passages, for instance, can tell you that students are reading below grade level, but they are unable to tell you *why.* Insofar as that is the case, they aren't particularly good for actually helping schools improve. Many states, consequently, were faced with the task of taking a school district into receivership before actually determining what they needed to do to strengthen local capacity.[55] Others simply closed schools down; in New York City alone, roughly 150 schools were shuttered for low performance between 2002 and 2014.

Twelve years later, no state had succeeded in moving all students to levels of proficiency, as required by the law. Though many individual schools had met their targets, and though test scores on the whole were up modestly from a decade earlier, the vast majority of schools had not met their goals, and achievement gaps continued to persist. Facing a potential crisis, the U.S. Department of Education began issuing waivers to states in 2011, freeing them from NCLB's accountability mechanisms. In return for such a reprieve, states were asked to adopt new standards for college readiness—usually the Common Core State Standards—and to tie teacher evaluations to student achievement data.[56] The Department of Education also urged states to propose new accountability frameworks that included factors like test score growth and graduation rates. Bargains were ultimately struck with the vast majority of states.[57]

Thus, roughly a decade after its passage, NCLB—the most sweeping federal intervention ever into education—came to a quiet end. Rather than bringing all schools to proficiency, the law produced only minor

—-1
—0
—+1

gains in student achievement—some of which have been linked to cheating, and many of which have been linked to the phenomenon of teaching to the test. A consensus emerged, even among many of the law's initial supporters, that it had failed.[58]

But the phaseout of NCLB did not mark the end of the standards and accountability era. Criticisms were commonplace. And even many of the law's early supporters faulted NCLB's unrealistic expectations, its sticks-instead-of-carrots approach, and its harsh punishments for low performers. But it had become abundantly clear that, whatever the opposition to high-stakes tests, the practice of holding schools accountable for measureable outcomes wasn't going away. Not by a long shot.

In December 2015, President Barack Obama signed the Every School Succeeds Act (ESSA), reauthorizing the ESEA, and replacing NCLB. But though ESSA eased up on the punitive features of NCLB, it did not wipe the slate clean. As a U.S. Department of Education overview put it: "NCLB put in place measures that exposed achievement gaps among traditionally underserved students and their peers and spurred an important national dialogue on education improvement. This focus on accountability has been critical in ensuring a quality education for all children, yet also revealed challenges in the effective implementation of this goal. Parents, educators, and elected officials across the country recognized that a strong, updated law was necessary to expand opportunity to all students; [to] support schools, teachers, and principals; and to strengthen our education system and economy."[59] Though ESSA allowed more flexibility with regard to testing, it still required states to test students in grades three through eight, as well as once in high school, for both math and English. Consequently, testing still rules the day. By the time he or she finishes high school, the average American student has sat through roughly ten standardized tests a year for at least seven years.[60]

What about All Those Other Tests?

Of course, state-mandated exams are not alone in the testing ecosystem. And though no other tests are used to, say, close a school down, the impact of an array of other exams can still be felt in policy discussions and parental decision making.

The oldest of these other tests is the SAT, developed in the 1920s by a consortium of prominent colleges and universities.[61] Although the test was not designed to measure school quality, SAT scores are often cited as evidence of student preparedness for college. By extension, SAT scores are commonly, if problematically, used to measure high school effectiveness, so it is perhaps worth briefly discussing the shortcomings of the SAT.

The first problem with the SAT is the problem of new information. Much of what is covered on the SAT is learned in school. On the surface this seems fine; yet colleges and universities already know a great deal from transcripts and grades about what students learned in elementary and high school. Such overlap might seem harmless until we consider factors like test anxiety, poor test-taking skills, or a bad night's sleep—all of which can affect a single day of testing in a way they wouldn't affect a student's grade in a yearlong course. That isn't to say that grades are perfect; they aren't. But it should raise questions about high-stakes uses of a test like the SAT.

The second problem facing the SAT is cultural and class bias. SAT vocabulary words, for instance, are more likely to be in common use in some households than in others. Specifically, white students from higher-income families are more likely to have been exposed to some of the arcane language used on the test. Additionally, they are more likely to have been drilled by their parents on "SAT words."[62]

A related problem facing the SAT is that of coachability. Companies like Kaplan and Princeton Review have long staked their livelihoods

—-1
—0
—+1

on a guarantee—that your score will go up after a four- to ten-week-long test-prep course. And, generally, scores *do* go up. That's great if you've invested time and money in an SAT boot camp; but it should raise questions about the degree to which the test tells us anything about innate ability or school learning. Higher scores may, it seems, tell us very little about those things, and far more about how many test-taking tricks a student has learned or how many practice tests he or she has taken. More disturbingly, the cost of such courses—generally around $500—inherently disadvantages low-income students.[63]

Perhaps the clearest and most cogent criticism of the SAT as a measure of school quality comes from a report by the College Board itself—the corporate parent of the SAT. According to a 1985 College Board report: "Those who reject the SAT as a barometer of schooling are on firm ground. Students who take the test are a representative sample of neither high school seniors nor college-bound seniors. The SAT was never intended to represent all of the important areas of understanding, knowledge, or skill—not to mention constructive attitudes, values, or other noncognitive characteristics—in which schools aim to bring about student growth. Moreover, the scores are not affected only by formal schooling; they measure abilities that are developed both in and out of school."[64] Basically, SAT scores are inappropriate for measuring K–12 education.

The ACT, the SAT's biggest rival, is a somewhat different test. Rather than seeking to measure a student's reasoning ability, the ACT seeks to measure achievement in English, math, science, reading, and writing. Given this aim, the ACT does not suffer from some of the weaknesses of the SAT—a fact that has prompted the College Board to adapt its test and make it more like that of its rival. Still, ACT scores provide little new information to colleges, can be significantly influenced by studying, and correlate highly with both student culture and family income.[65]

-1—
0—
+1—

Three more tests are worth discussing here—one designed to goad high school students into working harder, and two designed to measure the quality of the nation's schools.

The first of those is the high school exit exam—a test predicated on the belief that high school students do not work as hard as they could because gaining a diploma is too easy. Acting on that belief, roughly half of the states now administer such tests, and many require passing scores for graduation. Because such exams suffer from the same problems as other state-run standardized testing programs, however, they tend to disadvantage particular students. The result, as research has revealed, is that exit exams significantly reduce the probability of completing high school, particularly for low-income students and students of color. Insofar as such tests appear to exacerbate inequality, many view them as misguided.[66]

Among the two tests used to measure the quality of American schools, the older, and exclusively domestic, test is the National Assessment of Educational Progress (NAEP). Created in the 1960s to serve as "the nation's report card," NAEP tests are given to sample populations in grades four, eight, and twelve. The tests assess a wide range of subjects and are issued every year. The long-term trend assessments are typically administered every four years.

There are a few advantages to NAEP over the state-run achievement tests mandated by federal law. First, not every student has to take the NAEP, meaning that the instructional time lost to testing is minimized. Statistically speaking, there's no need to test every student if a representative sample can be assembled—and that's exactly what NAEP does. Second, because schools and districts are not held accountable for NAEP scores, there is very little likelihood of anyone teaching to the test, and no pressure is directed at children to improve their NAEP scores. Third, NAEP tests a broader range of school subjects than high-stakes achievement tests do, meaning that even if

NAEP *were* used for accountability purposes, it would be less likely to narrow the curriculum.

Of course, NAEP has its limitations. NAEP is still a standardized achievement test that fails to account for prior knowledge and that relies on machine scoring. Additionally, because student academic achievement is the product of a student's out-of-school experiences even more than of his or her in-school experiences, NAEP scores ultimately may tell us more about factors like student poverty than they do about a school's quality. Finally, while NAEP may tell us something about academic achievement, it fails to tell us much else about schools.

The second test of national education outcomes is the Programme for International Student Assessment (PISA)—an exam created by the Organisation for Economic Cooperation and Development and given every three years to fifteen-year-olds from across the globe. PISA has many of the same advantages and disadvantages as NAEP. Given those disadvantages, it means that countries with lower levels of poverty, or stronger orientation toward test preparation, will outscore the United States. These score discrepancies have caused no small degree of handwringing. At the top of the annual heap are the usual suspects from East Asia: Singapore, Taiwan, South Korea, and three Chinese cities—Shanghai, Hong Kong, and Macao. The United States usually comes in somewhere in the middle of the pack, which usually elicits a reaction like that of a 2013 *New York Times* editorial: that America's students are falling "further and further behind."[67]

Even if we were to assume that PISA is a gold-standard measure, we might take heart in the fact that there is no statistically significant difference between the performance of American students and those in a place like Norway, which in 2015 was ranked by the United Nations as the best place in the world to live.

But tests like PISA should raise serious questions among observers. How can a test measure all of the different things that students learn

across various national school systems? After all, comparative measures require a common denominator, and no such denominator currently exists. PISA's approach is to measure a set of "real-world" skills unrelated to national curricula. This may allow for comparison, but it creates an obvious problem: the test measures what many schools *do not teach.* It establishes a level playing field, but it does so by ignoring what each system is actually trying to accomplish. "This," as my colleague David Labaree put it, "is leveling the playing field with a bulldozer." These issues are further compounded by the fact that the validity of PISA test items is unsupported by research. All of these issues make the most common use of PISA—decrying the state of public education—at best inappropriate and at worst absurd.[68] Whatever the interpretations of journalists and policy leaders, none of these tests indicates much about the quality of American schools.

Caveat Emptor

Tests usually do measure something about what a student knows. And when implemented thoughtfully, feedback from testing can provide information to educators and policymakers that may help them engage in systemic planning. It can also help students understand how they are doing relative to their peers and relative to expectations, at least on the subject area being tested.

But tests are incomplete measures of school quality. In fact, they are incomplete measures of student academic achievement, and academic achievement constitutes only a single component of a good school. How else can we explain the research finding that high schools effective at improving test scores are not necessarily effective at reducing dropout rates? Or the research finding that schools can have a major impact on students' lives—leading to lower likelihoods of arrest and higher rates of college attendance—without raising test scores?[69]

—-1
—0
—+1

Good schools do many things. They are places where children learn about the world and begin to imagine life beyond their neighborhoods. They are places where the arts are valued and pursued—where children learn to draw and dance and play the piano, as well as to understand a poem or a painting or a piece of music. They are places where ideas are sought and explored—for the purpose of expanding young people's notions of justice, broadening their visions of the possible, and welcoming them into ongoing cultural conversations. Our best schools are places where children gain confidence in themselves, build healthy relationships, and develop values congruent with their own self-interest. They are places of play and laughter and discovery.

Of course, good schools also promote student learning in core content areas, but measuring something as complicated as student learning, it turns out, is particularly hard when it has to be done in a uniform and cost-restricted way. It is particularly challenging given the fact that all students come to school with different prior levels of ability and preparation.

Tests have long been at the core of the American educational system. Consequently, most stakeholders view tests with less skepticism than they perhaps should. In one recent poll, for instance, more than two-thirds of respondents expressed support for federally required annual testing. Yet the public is increasingly bristling under the testing regime. Roughly two-thirds of respondents to the 2015 Phi Delta Kappa / Gallup poll expressed a concern that children are subject to too many tests, and they deemed test scores the least accurate of currently available measures of school effectiveness. There is, it seems, an expanding rift between Americans' historically rooted acceptance of testing and a rising uneasiness about the uses to which test scores are being put.[70]

Still, the public often lacks the tools to critique the testing regime. In a 2016 letter to The Ethicist column in the *New York Times Maga-*

zine, for example, a parent asked for advice about where to send his child. "State test scores came out recently," he wrote, "and our neighborhood public school, which is filled with some of the city's poorest kids, scored very low." Despite this parent's implicit recognition of the link between poverty and test scores, and despite the fact that he volunteers at the school and finds it "perfectly fine," he nevertheless concluded that there must be "something seriously wrong with how the school is educating kids."[71] That is certainly a possibility. In fact, it is a possibility at all schools, regardless of test scores. But what if the school, as this parent's intuition and experience tell him, is actually "perfectly fine"? What if the data available, in the form of standardized test scores, present a distorted picture of reality?

—-1
—0
—+1